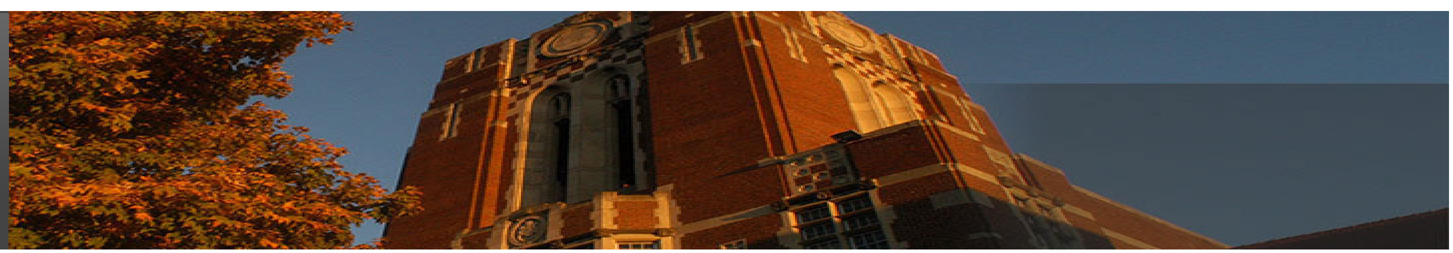
SHIP: Scalable Hierarchical Power Control for Large-Scale Data Centers

Xiaorui Wang, **Ming Chen**
University of Tennessee, Knoxville, TN

Charles Lefurgy, Tom W. Keller
IBM Research, Austin, TX

THE UNIVERSITY of TENNESSEE UT
KNOXVILLE



Introduction

- Data centers are expanding to meet new business requirement.
  - Cost-prohibitive to expand the power facility.
  - Upgrades of power/cooling systems lag far behind.
  - Example: NSA data center

- Power overload may cause system failures.
  - Power provisioning CANNOT guarantee exempt of overload.
  - Over-provisioning may cause unnecessary expenses.

Power control for an entire data center is very necessary.

THE UNIVERSITY of TENNESSEE UT KNOXVILLE

Challenges

- Scalability: One centralized controller for thousands of servers?
- Coordination: if multiple controllers designed, how do they interact with each other?
- Stability and accuracy: workload is time-varying and unpredictable.
- Performance: how to allocate power budgets among different servers, racks, etc.?

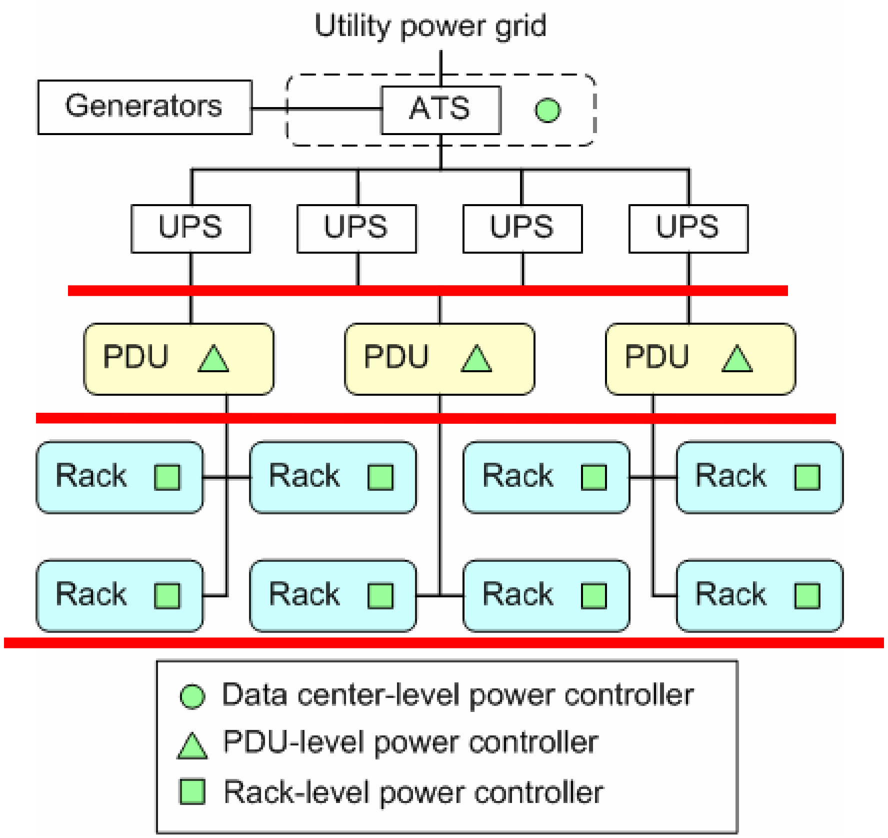
State of The Art

- Reduce power by improving energy-efficiency : [Lefurgy], [Nathuji], [Zeng], [Lu], [Brooks]
  - Based on heuristic and NOT enforce power budget.
- Power control for a server [Lefurgy], [Skadron], [Minerick], a rack, [Wang], [Ranganathan], [Femal]
  - Cannot be directly applied for data centers.
- *No "Power" Struggles* presents a multi-level power manager. [Raghavendra]
  - NOT designed based on power supply hierarchy
  - NO rigorous overall stability analysis
  - Only simulation results for 180 servers
- Use power as a knob to control performance requirements in OS level. [Horvath], [Chen], [Sharma]

What is This Paper About?

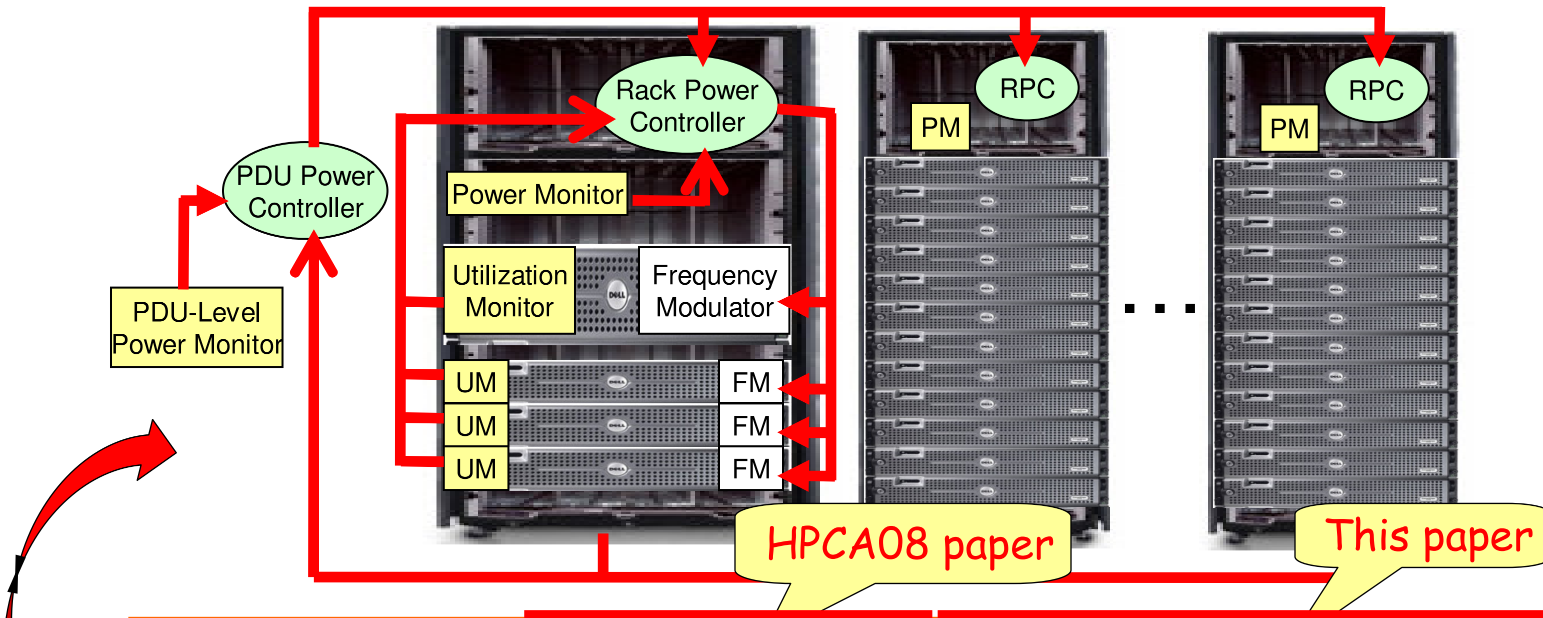
- SHIP: a highly Scalable Hierarchical Power control architecture for large-scale data centers
  - Scalability: decompose the power control for a data center into three levels.
  - Coordination: hierarchy is based on power distribution system in data centers.
  - Stability and accuracy: theoretically guaranteed by Model Predicative Control (MPC) theory.
  - Performance: differentiate power budget based on performance demands, *i.e.* utilization.

Power Distribution Hierarchy



- A simplified example for a three-level data center
  - Data center-level
  - PDU-level
  - Rack-level
- Thousands of servers in total

Control Architecture



	Rack-level	PDU-level	Data center-level
Controlled variable	The total power of the rack	The total power of the PDU	The total power of the data center
Manipulated variable	The CPU frequency of each server	The power budget of each rack	The power budget of each PDU

THE UNIVERSITY of TENNESSEE UT KNOXVILLE

PDU-level Power Model

- System model:

$$pp(k+1) = pp(k) + \sum_{i=1}^{N} \Delta pr_i(k)$$

$pp(k)$: the total power of PDU

$\Delta pr_i(k)$: the power change of rack *i*

- Uncertainties:

$$\Delta pr_i(k) = g_i \Delta br_i(k)$$

$\Delta br_i(k)$: the change of power budget for rack *i*

g_i is the power change ratio.

- Actual model:

$$pp(k+1) = pp(k) + [g_1 \quad \dots \quad g_N] \begin{bmatrix} \Delta br_1(k) \\ \dots \\ \Delta br_N(k) \end{bmatrix}$$

Model Predictive Control (MPC)

- Control objective:

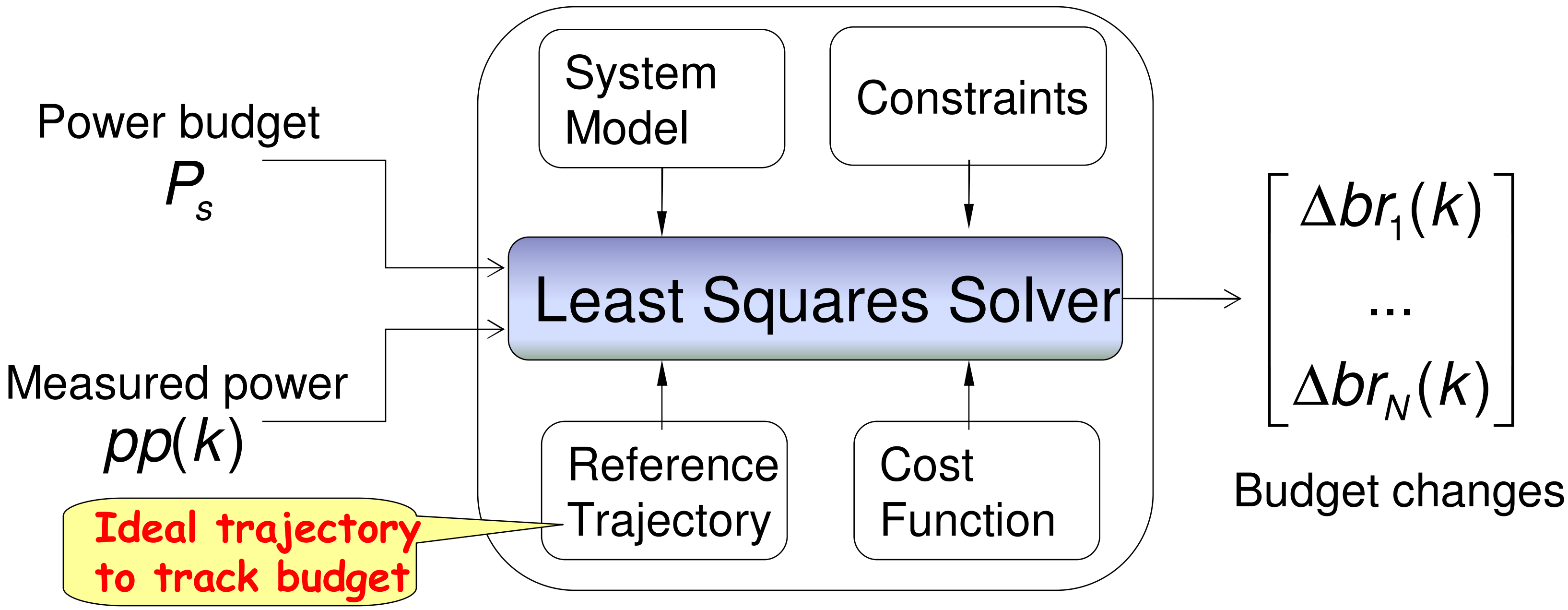
$$\min_{\{\Delta br_j(k)|1\le j\le N\}} (pp(k+1)-P_s)^2$$

$$subject\ to: P_{\min,j} \le \Delta br_j(k) + br_j(k) \le P_{\max,j} (1\le j\le N)$$

$$pp(k+1) \le P_s$$

- Design steps:
  - Design a dynamic model for the controlled system.
  - Design the controller.
  - Analyze the stability and accuracy.

MPC Controller Design



$$V(k) = \sum_{i=1}^{P} \| pp(k+i \mid k) - ref(k+i \mid k) \|_{Q(i)}^{2} + \sum_{i=0}^{M-1} \| \mathbf{\Delta br(k+i \mid k) + br(k+i \mid k) - P_{max}} \|_{\mathbf{R(i)}}^{2}$$

Tracking error — Control penalty

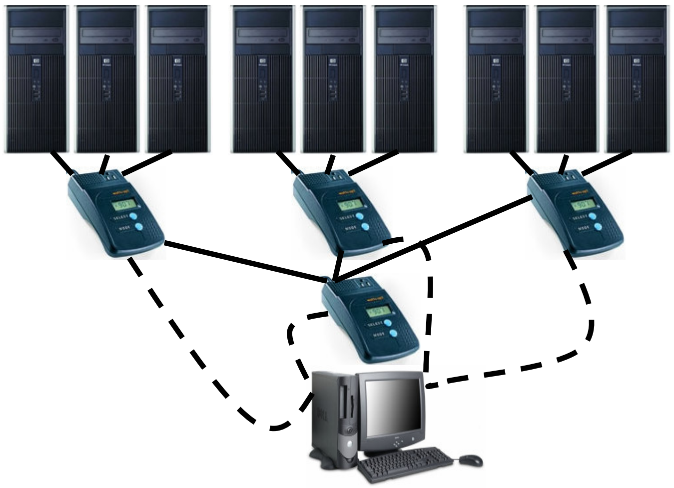
THE UNIVERSITY of TENNESSEE UT KNOXVILLE

Stability

- Local Stability
  - g_i is assumed to be 1 at design time.
  - g_i is unknown a priori.
  - $0 < g_i < 14.8$: 14.8 times of the allocated budget
- Global Stability
  - Decouple controllers at different levels by running them in different time scales.
  - The period of upper-level control loop > the settling time of the lower-level
  - Sufficient but not necessary

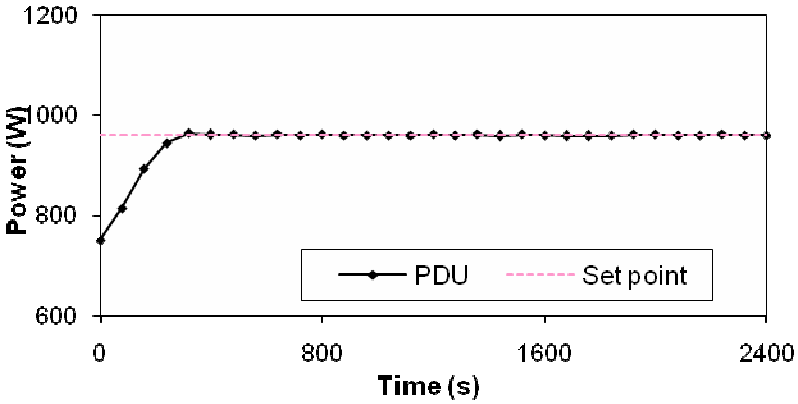
System Implementation

- Physical testbed
  - 10 Linux servers
  - Power meter (Wattsup)
    - error: ±1.5%
    - sampling period: 1 sec
  - Workload: HPL, SPEC
  - Controllers:
    - call matlab function.
    - period: 5s for rack, 30s for PDU

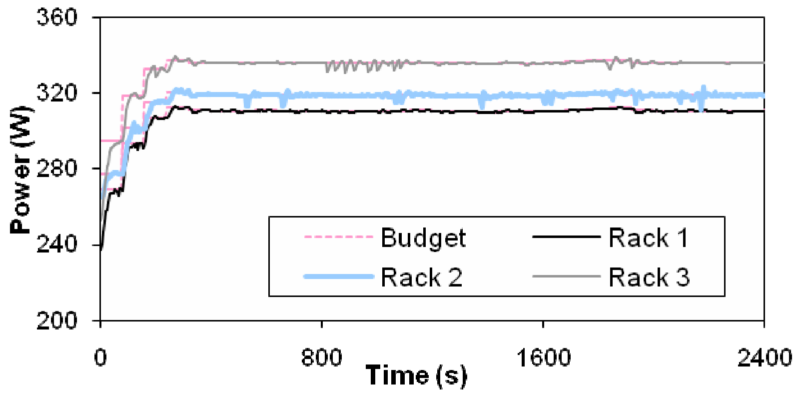


- Simulator (C++)
  - Simulate large-scale data centers in three levels.
  - Utilization trace file from 5,415 servers in real data centers
  - Power model is based on experiments in servers.
  - Generate 3 data center configurations.

Precise Power Control (Testbed)



- Power can be precisely controlled at the budget.
- The budget can be reached within 4 control periods.

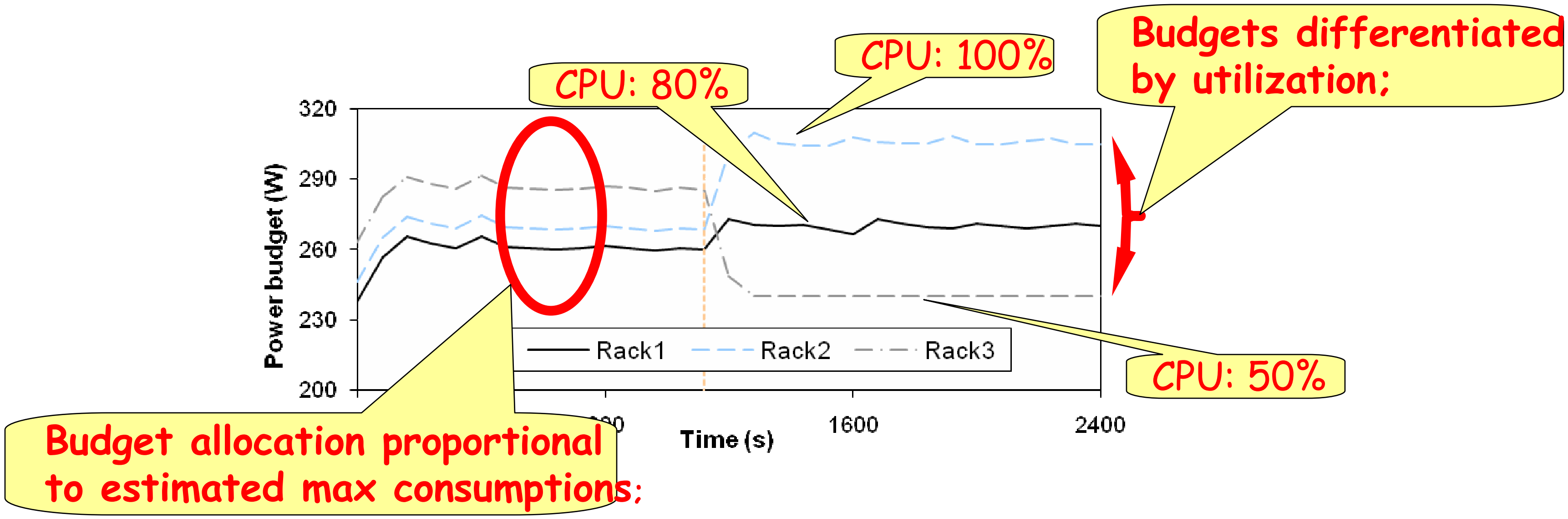


- The power of each rack is controlled at their budgets.
- Budgets are proportional to P_{max} .

- Tested under other set points

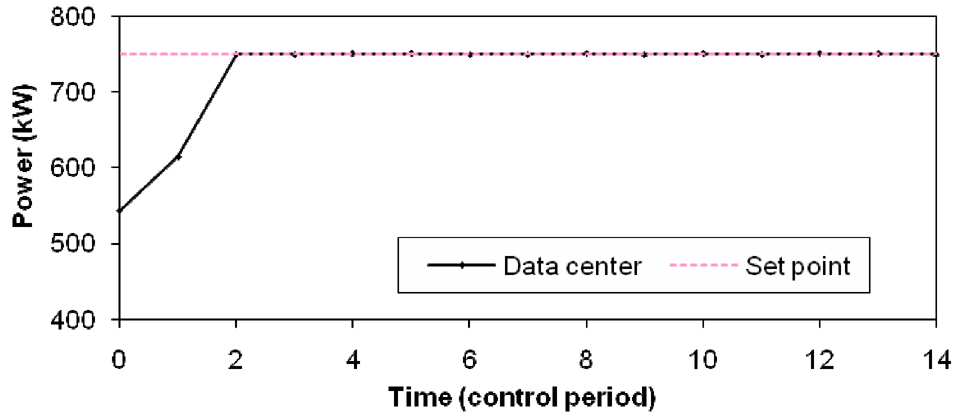
THE UNIVERSITY of TENNESSEE UT KNOXVILLE

Power Differentiation (Testbed)

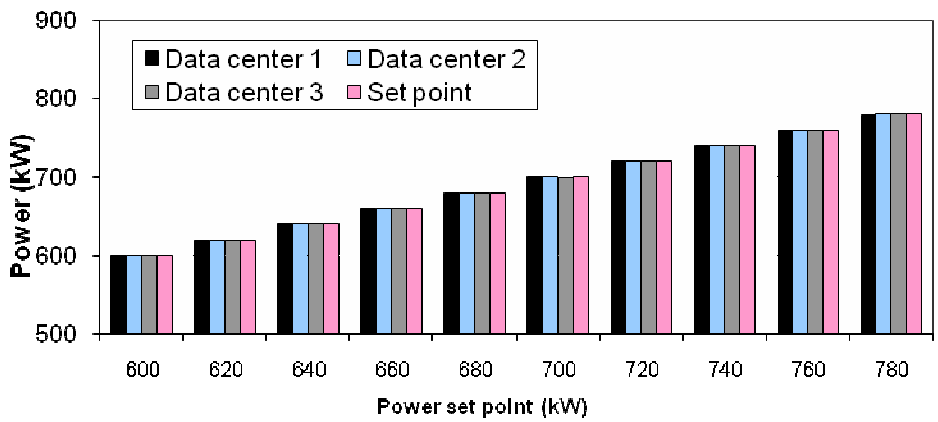


- Capability to differentiate budgets based on workload to improve performance
- Take the utilization as the optimization weights.
- Other differentiation metrics: response time, throughput

Simulation for Large-scale Data Centers



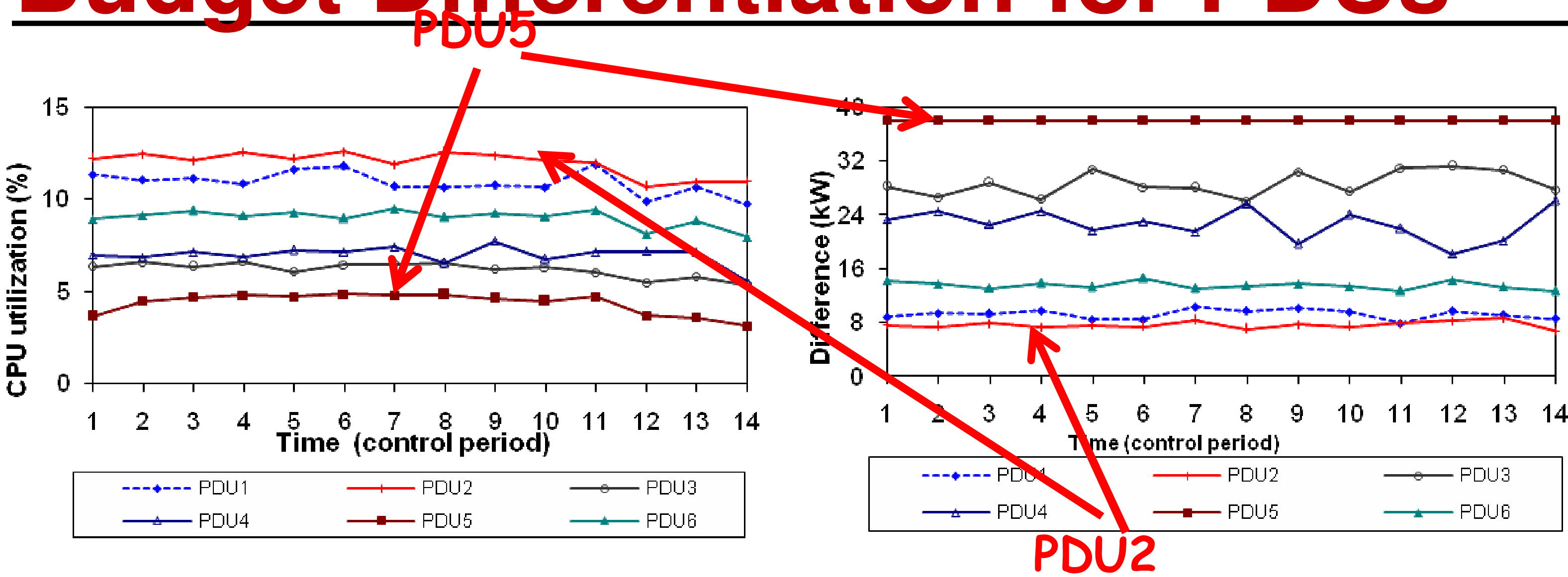
- 6 PDU, 270 racks
- Real data traces
- 750 kW



- Randomly generate 3 data centers
- Real data traces

THE UNIVERSITY of TENNESSEE UT KNOXVILLE

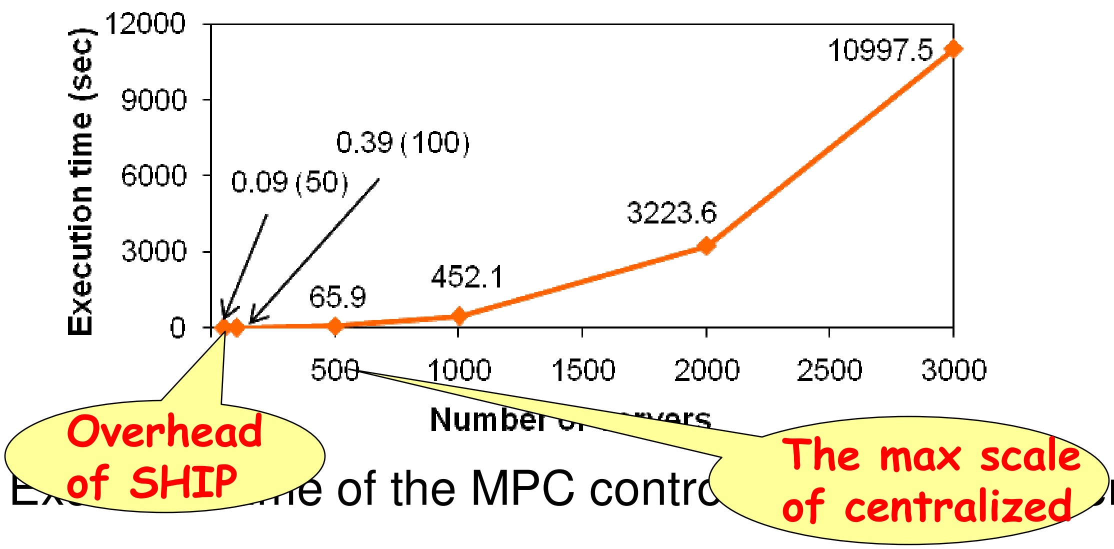
Budget Differentiation for PDUs



- Power differentiation in large-scale data centers;
  - Minimize the difference with estimated max power consumption.
  - Utilization is the weight.
  - The difference order is consistent with the utilization order.

THE UNIVERSITY of TENNESSEE UT KNOXVILLE

Scalability of SHIP



Ex... of the MPC contr... ...rvers

	Centralized	SHIP
Level	One level	Multiple
Computation overhead	Large	Small
Communication overhead	Long	Short
Scalability	NO	YES

Conclusion

- SHIP: a highly Scalable HIerarchical Power control architecture for large-scale data centers
  - Three-levels: rack, PDU, and data center
  - MIMO controllers based on optimal control theory (MPC)
  - Theoretically guaranteed stability and accuracy
  - Discussion on coordination among controllers
- Experiments on a physical testbed and a simulator
  - Precise power control
  - Budget differentiation
  - Scalable for large-scale data centers

Acknowledgment

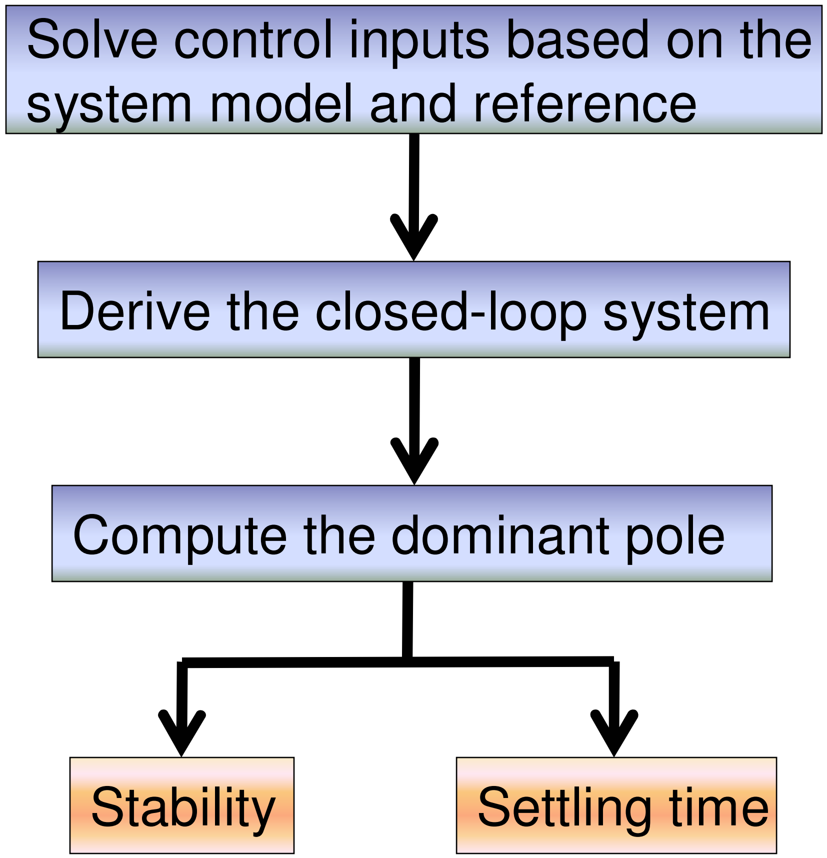
- This work was supported, in part, by
  - NSF under a CAREER Award CNS-0845390 and a CSR grant CNS-0720663
  - Microsoft Research under a power-aware computing award in 2008

Thank you!

THE UNIVERSITY of TENNESSEE UT KNOXVILLE

Backup Slides

Stability Analysis



More Implementation Details

- CPU modulator
  - 4-5 frequency levels to scale
  - fraction levels:
    - For 2.8, that is: 2, 3, 3, 3, 3 with 5 subintervals.
  - 50 subintervals in each period of rack controllers
- Trace file
  - From 5415 servers in multiple data centers (manufacturing, financial, telecommunication, retail sectors)
  - Average CPU utilization every 15 minutes
  - From 00:00 on July 14 to 23:45 on July 20 in 2008

Reference Trajectory

$$ref(k+i \mid k) = P_s - e^{-\frac{T_p}{T_{ref}} i}(P_s - pp(k)),\ 1 \le i \le P$$

- T_p and T_{ref} specify the speed of system response.
- P: prediction horizon

THE UNIVERSITY of TENNESSEE UT KNOXVILLE